The Future Evolution of High-Performance Microprocessors

Norm Jouppi
HP Labs

© 2005 Hewlett-Packard Development Company, L.P.
The information contained herein is subject to change without notice

Keynote Overview

- What, Why, How, and When of Evolution
- Microprocessor Environmental Constraints
- The Power Wall
- Power: From the Transistor to the Data Center
- The Evolution of Computer Architecture Ideas
- Summary

Disclaimer

- These views are mine, not necessarily HP
- "Never make forecasts, especially about the future" – Samuel Goldwyn

What is Evolution

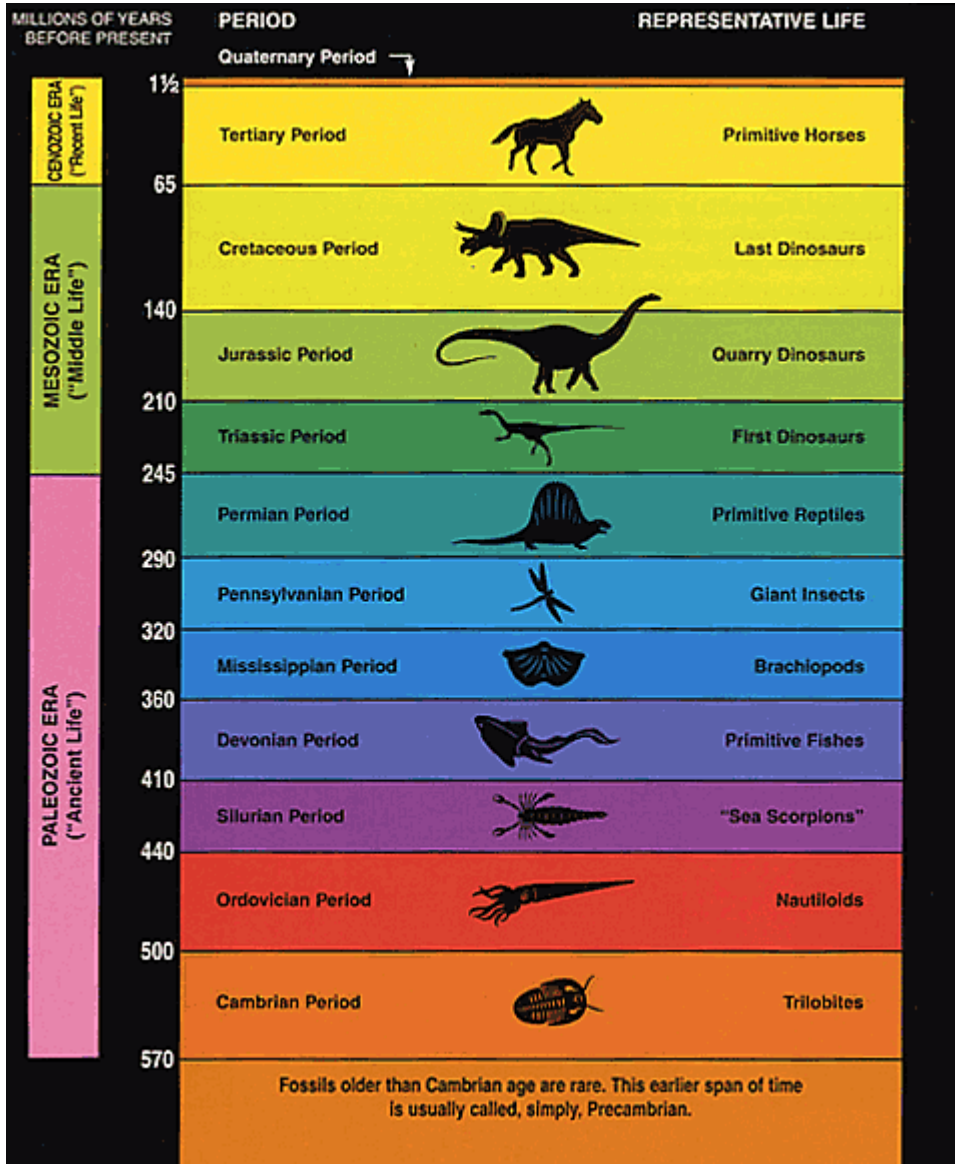
- Definition:
  - a process of *continuous* change from a lower, simpler, or worse to a higher, more complex, or better state
  - unfolding

Why Evolution

- Evolution is a Very Efficient Means of Building New Things
  - Reuse, recycle
  - Minimum of new stuff
  - Much easier than revolution

When Evolution

- Can be categorized into Eras, Periods, etc.



Technology

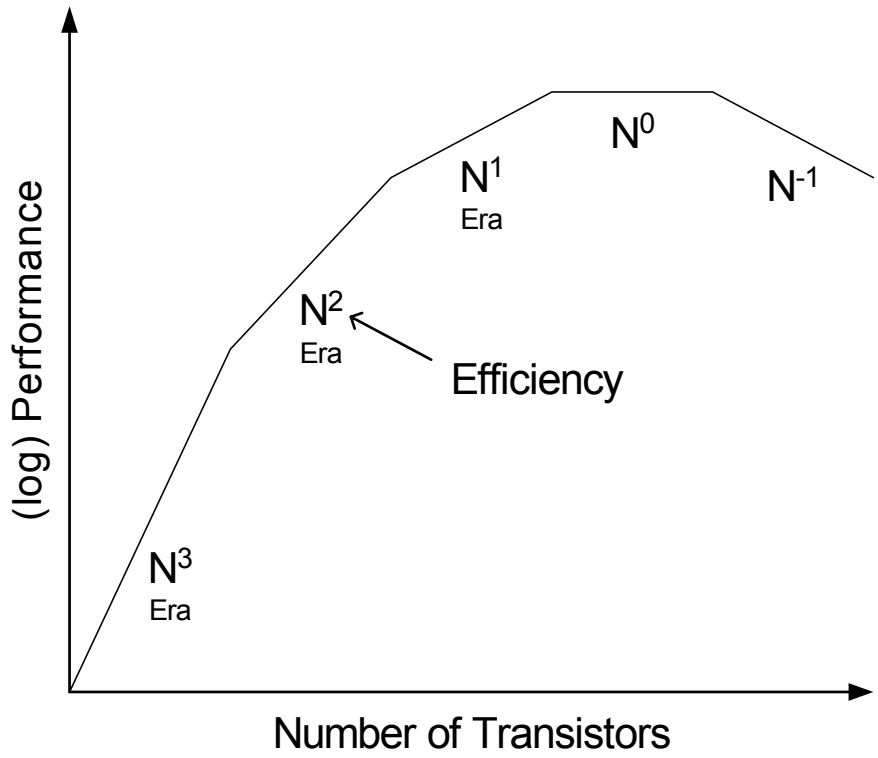
- Usually evolution, not revolution
- Many revolutionary technologies have a bad history:
  - Bubble memories
  - Josephson junctions
  - Anything but Ethernet
  - Etc.
- Moore's Law has been a key force driving evolution of the technology

Moore's Law

- Originally presented in 1965
- Number of transistors per chip is $1.59^{\text{year-1959}}$ (originally $2^{\text{year-1959}}$)
- Classical scaling theory (Denard, 1974)
  - With every feature size scaling of n
    - You get $O(n^2)$ transistors
    - They run O(n) times faster
- Subsequently proposed:
  - "Moore's Design Law" (Law #2)
  - "Moore's Fab Law" (Law #3)

Microprocessor Efficiency Eras (Jouppi)

- Moore's Law says number of transistors scaling as $O(n^2)$ and speed as $O(n)$
- Microprocessor performance should scale as $O(n^3)$



N^3 Era

- n from device speed, n^2 from transistor count
- 4004 to 386
- Expansion of data path widths from 4 to 32 bits
- Basic pipelining
- Hardware support for complex ops (FP mul)
- Memory range and virtual memory
- Hard to measure performance
  - Measured in MIPS

N^3 Era

- n from device speed, n^2 from transistor count
- 4004 to 386
- Expansion of data path widths from 4 to 32 bits
- Basic pipelining
- Hardware support for complex ops (FP mul)
- Memory range and virtual memory
- Hard to measure performance
  - Measured in MIPS
  - But how many 4-bit ops = 64-bit FP mul?

N^3 Era

- n from device speed, n^2 from transistor count
- 4004 to 386
- Expansion of data path widths from 4 to 32 bits
- Basic pipelining
- Hardware support for complex ops (FP mul)
- Memory range and virtual memory
- Hard to measure performance
  - Measured in MIPS
  - But how many 4-bit ops = 64-bit FP mul?
    - More than 1500!

N^2 Era

- n from device speed, only n from n^2 transistors
- 486 through Pentium III/IV
- Era of large on-chip caches
  - Miss rate halves per quadrupling of cache size
- Superscalar issue
  - 2X performance from quad issue?
- Superpipelining
  - Diminishing returns

N Era

- n from device frequency, 1 from n^2 transistor count
- Very wide issue machines
  - Little help to many aps
  - Need SMT to justify
- Increase complexity and size too much → slowdown
  - Long global wires
  - Structure access times go up
  - Time to market

Environmental Constraints on Microprocessor Evolution

- Several categories:
- Technology scaling
  - Economics
  - Devices
  - Voltage scaling
- System-level constraints
  - Power

Supply Economics: Moore's Fab Law

- Fab cost is scaling as 1/feature size
- 90nm fabs currently cost 1-2 billion dollars
- Few can afford one by themselves (except Intel)
  - Fabless startups
  - Fab partnerships
    - IBM/Toshiba/etc.
    - Large foundries
- But number of transistors scales as 1/feature size^2
  - Transistors still getting cheaper
  - Transistors still getting faster

Supply Economics: Moore's Design Law

- The number of designers goes as O(1/feature)
- The 4004 had 3 designers, 10μm
- Recent 90nm microprocessors have ~300 designers
- Implication: design cost becomes very large
- Consolidation in # of viable microprocessors
- Microprocessor cores often reused
  - Too much work to design from scratch
  - But "shrinks and tweaks" becoming difficult in deep submicron technologies

Devices

- Transistors historically get faster $\propto$ feature size
- But transistors getting much leakier
  - Gate leakage (fix with high-K gate dielectrics)
  - Channel leakage (dual-gate or vertical transistors?)
- Even CMOS has significant static power
  - Power is roughly proportional to # transistors
  - Static power approaching dynamic power
  - Static power increases with chip temperature
    - Positive feedback is bad

Voltage Scaling

- High performance MOS started out with12V
- Max Voltage scaling roughly as sqrt(feature)
- Power is $\propto CV^2f$
  - Lower voltage can reduce the power as square
  - But speed goes down with lower voltage
- Current high-performance microprocessors have 1.1V supplies
- Reduced power $(12/1.1)^2 = 119X$ over 24 years!

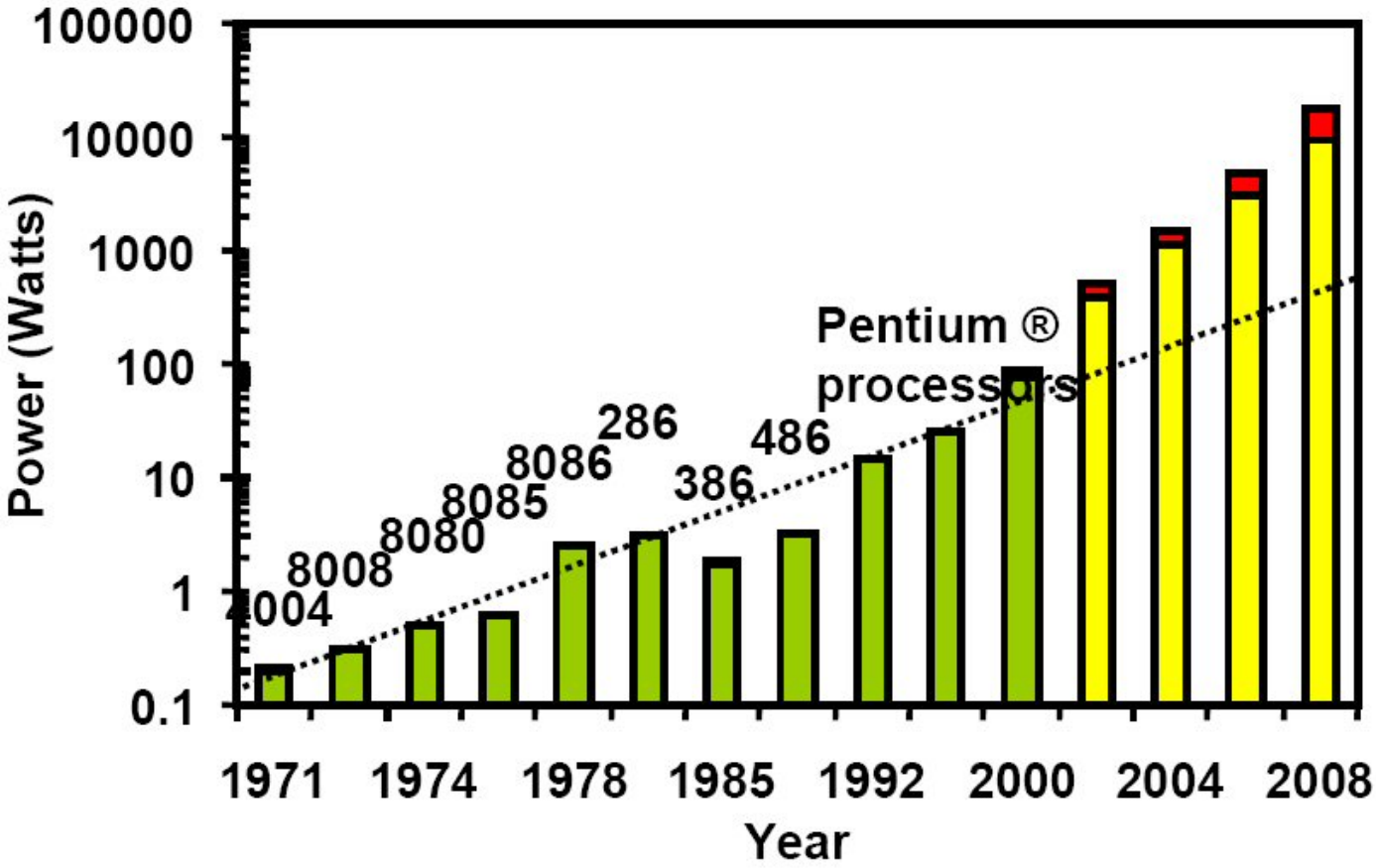
Limits of Voltage Scaling

- Beyond a certain voltage transistors don't turn off
- ITRS projects minimum voltage of 0.7V in 2018
  - Limited by threshold voltage variation, etc.
  - But high-performance microprocessors are now 1.1V
- Only $(1.1/0.7)^2 = 2.5X$ reduction left in next 14 years!

System-Level Power

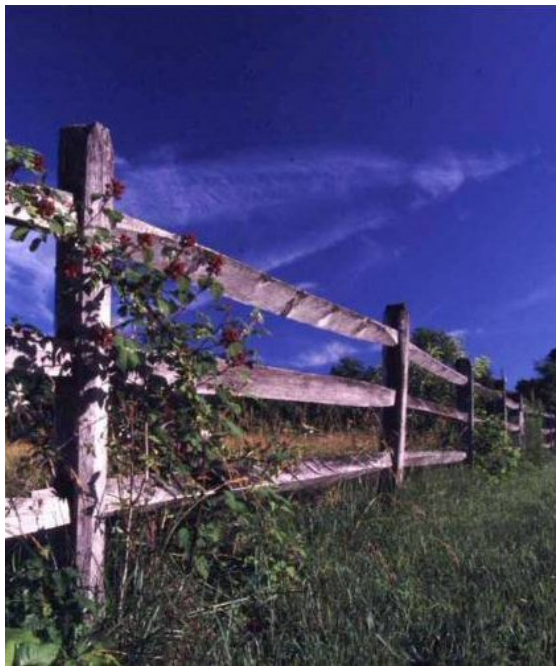
- Per-chip power envelope is peaking
  - Itanium 2 @ 130W => Montecito @ 100W
- 1U servers and blades reduces heat sink height
- Cost of power in and heat out for several years can equal original system cost
- First class design constraint

Current Microprocessor Power Trends



- Figure source: Shekhar Borkar, "Low Power Design Challenges for the Decade", Proceedings of the 2001 Conference on Asia South Pacific Design Automation, IEEE.

The Power Wall in Perspective



The Memory Wall

The Power Wall

Pushing Out the Power Wall

- "Waste Not Want Not" (Ben Franklin) for circuits
- Power-efficient microarchitecture
- Single threaded vs. throughput tradeoff

"Waste Not Want Not" for Circuits

- Lots of circuit ways to save power already in use
  - Clock gating
  - Multiple transistor thresholds
  - Sleeper transistors
  - Etc.
- Thus circuit-level power dissipation is already fairly efficient
- What about architectural savings?

Power-Efficient Microarchitecture

- Off-chip memory reference costs a lot of power
  - Thus drive to more highly-associative on-chip caches
  - Limits to how far this can go
- Lots of other similar examples
- PACS workshop/conference proceedings
  - No silver bullet
  - Limited benefits from each technique

Single-Threaded / Throughput Tradeoff

- Reducing transistors/core can yield higher MIPS/W
- Move back towards N^3 scaling efficiency
- Thus, expect trend to simpler processors
  - Narrower issue width
  - Shallower pipelines
  - More in-order processors or smaller OOO windows
- "Back to the Future"
- But this gives lower single-thread performance
  - Can't simplify core too quickly
- Tradeoffs on what to eliminate not always obvious
  - Examples: Speculation, Multithreading

Speculation

- Is speculation uniformly bad?
  - No
- Example: branch prediction
  - Executing down wrong path wastes performance & power
  - Stalling at every branch would hurt performance & power
    - Circuits leak when not switching
- Predicting a branch can save power
  - Plus predictor memory takes less power/area than logic
- But current amount of speculation seems excessive

Multithreading

- SMT is very useful in wide-issue OOO machines
  - Good news: increases power efficiency
  - Bad news: Wide issue still power inefficient
- Multithreading useful even in simple machines
  - During cache misses transistors still leak
    - Not enough time to gate power
  - May only need 2 or 3 thread non-simultaneous MT

Microarchitectural Research Implications

- Processors need to get simpler (not more complicated) to become more efficient
- More complicated microarchitecture mechanisms with little benefit not needed

Recap: Possible Future Evolution in Terms of $P = CV^2f$

- Formula doesn't change, but terms do:
  - Power — 10^0
    - Already at system limits, better if lower
  - Voltage — 10^0
    - Only a factor of 2.5 left over next 14 years
  - Clock Frequency — 10^0
    - Not scaling with device speed
    - Fewer pipestages, higher efficiency
    - Move from 30 to 10 stages from 90nm to 32 nm
  - Capacitance (~Area) needs to be repartitioned for higher system performance

Number of Cores Per Die

- Scale processor complexity as 1/feature
  - Number of cores will go up dramatically as n^3
  - From 1 core at 90nm to 27 per die at 30nm!
  - But can we efficiently use that many cores?

The Coming Golden Age

- We are on the cusp of the golden age of parallel programming
  - Every major application needs to become parallel
  - "Necessity is the mother of invention"
- How to make use of many processors effectively?
  - Some aps inherently parallel (Web, CFD, TPC-C)
  - Some applications are very hard to parallelize
  - Parallel programming is a hard problem
- Can't use large amounts of speculation in software
  - Just moves power inefficiency to a higher level

Important Architecture Research Problems

Not an exhaustive list:

- How to wire up CMPs:
  - Memory hierarchy?
  - Transactional memory?
  - Interconnection networks?
- How to build cores:
  - Heterogeneous CMP design
  - Conjoined core CMPs
- Power: From the transistor through the datacenter

Important Architecture Research #2

Many **system level** problems:

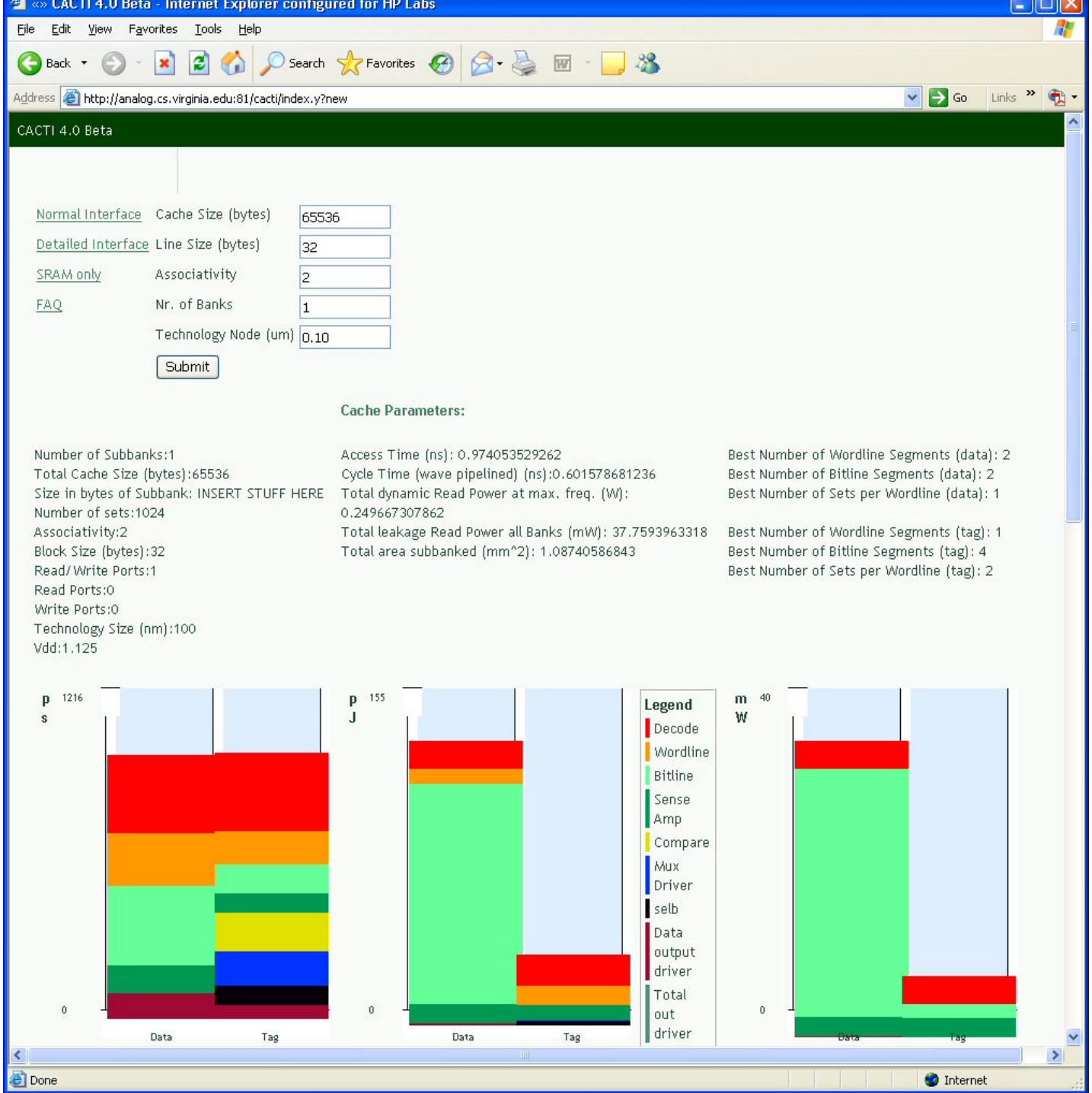
- Cluster interconnects
- Manageability
- Availability
- Security
- Hardware/software tradeoffs (ASPLOS) increasingly important

Power:
From the Transistor through the Data Center

- CACTI 4.0
- Heterogenous CMP
- Conjoined cores
- Data center power management

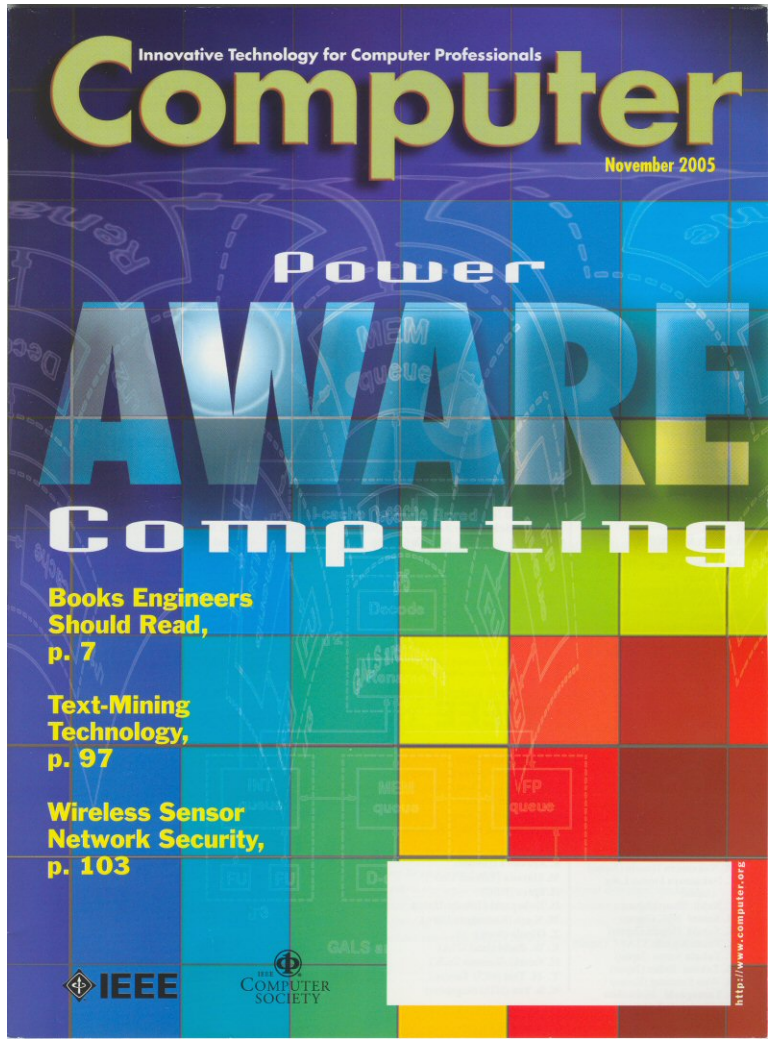
CACTI 4.0

- Collaboration with David Tarjan of UVa
- Now with leakage power model
- Also includes:
  - Much improved technology scaling
  - Circuit updates
  - Scaling to much larger cache sizes
  - Options: serial tag/data, high speed access, etc.
  - Parameterized data and tag widths (including no tags)
  - Beta version web interface: http://analog.cs.virginia.edu:81/cacti/index.y
  - Full release coming soon (norm.jouppi@hp.com)


«» CACTI 4.0 Beta - Internet Explorer configured for HP Labs
File Edit View Favorites Tools Help
Back Search Favorites
Address http://analog.cs.virginia.edu:81/cacti/index.y?new
Go Links
CACTI 4.0 Beta
Normal Interface
Detailed Interface
SRAM only
FAQ
Cache Size (bytes) 65536
Line Size (bytes) 32
Associativity 2
Nr. of Banks 1
Technology Node (um) 0.10
Submit
Cache Parameters:
Number of Subbanks:1
Total Cache Size (bytes):65536
Size in bytes of Subbank: INSERT STUFF HERE
Number of sets:1024
Associativity:2
Block Size (bytes):32
Read/Write Ports:1
Read Ports:0
Write Ports:0
Technology Size (nm):100
Vdd:1.125
Access Time (ns): 0.974053529262
Cycle Time (wave pipelined) (ns):0.601578681236
Total dynamic Read Power at max. freq. (W): 0.249667307862
Total leakage Read Power all Banks (mW): 37.7593963318
Total area subbanked (mm^2): 1.08740586843
Best Number of Wordline Segments (data): 2
Best Number of Bitline Segments (data): 2
Best Number of Sets per Wordline (data): 1
Best Number of Wordline Segments (tag): 1
Best Number of Bitline Segments (tag): 4
Best Number of Sets per Wordline (tag): 2
p s 1216
0
Data
Tag
p J 155
0
Data
Tag
Legend
Decode
Wordline
Bitline
Sense Amp
Compare
Mux Driver
selb
Data output driver
Total out driver
m W 40
0
Data
Tag
Done
Internet

Heterogeneous Chip Multiprocessors

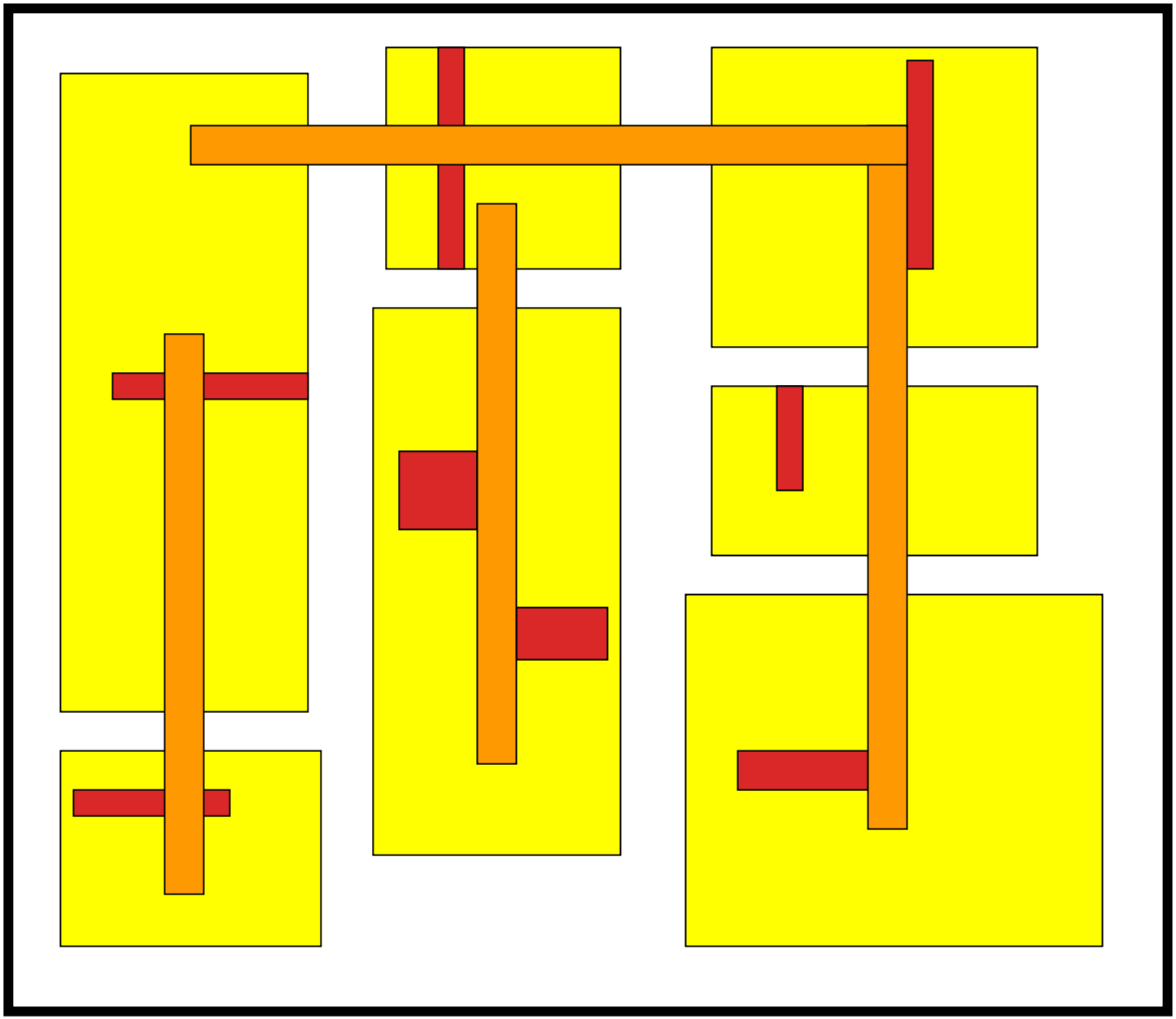
- A.k.a. Asymmetric, Non-homogeneous, synergistic,…
- Single ISA vs. Multiple ISA
- Many benefits:
  - Power
  - Throughput
  - Mitigating Amdahl's Law
- Open questions
  - Best mix of heterogeneity



Potential Power Benefits

- Grochowski et. al. ICCD 2004:
  - Asymmetric CMP => 4-6X
  - Further voltage scaling => 2-4X
  - More gating => 2X
  - Controlling speculation =>1.6X

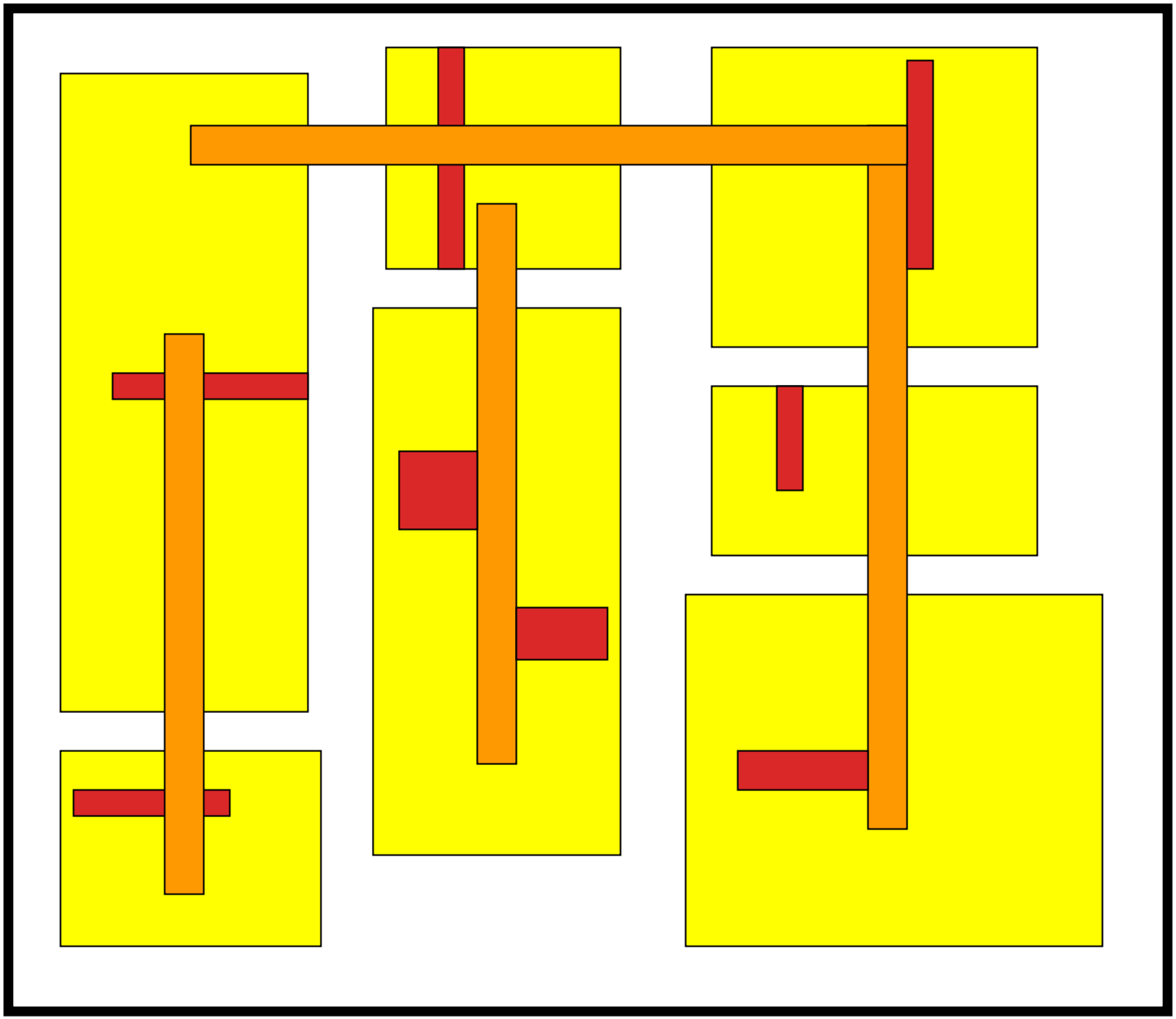
Clock Gating



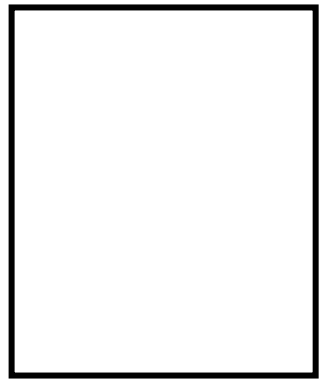
Transistors still leak

And busses are still long, slow, & waste power

Key: Yellow are inactive, red are active, orange are global busses

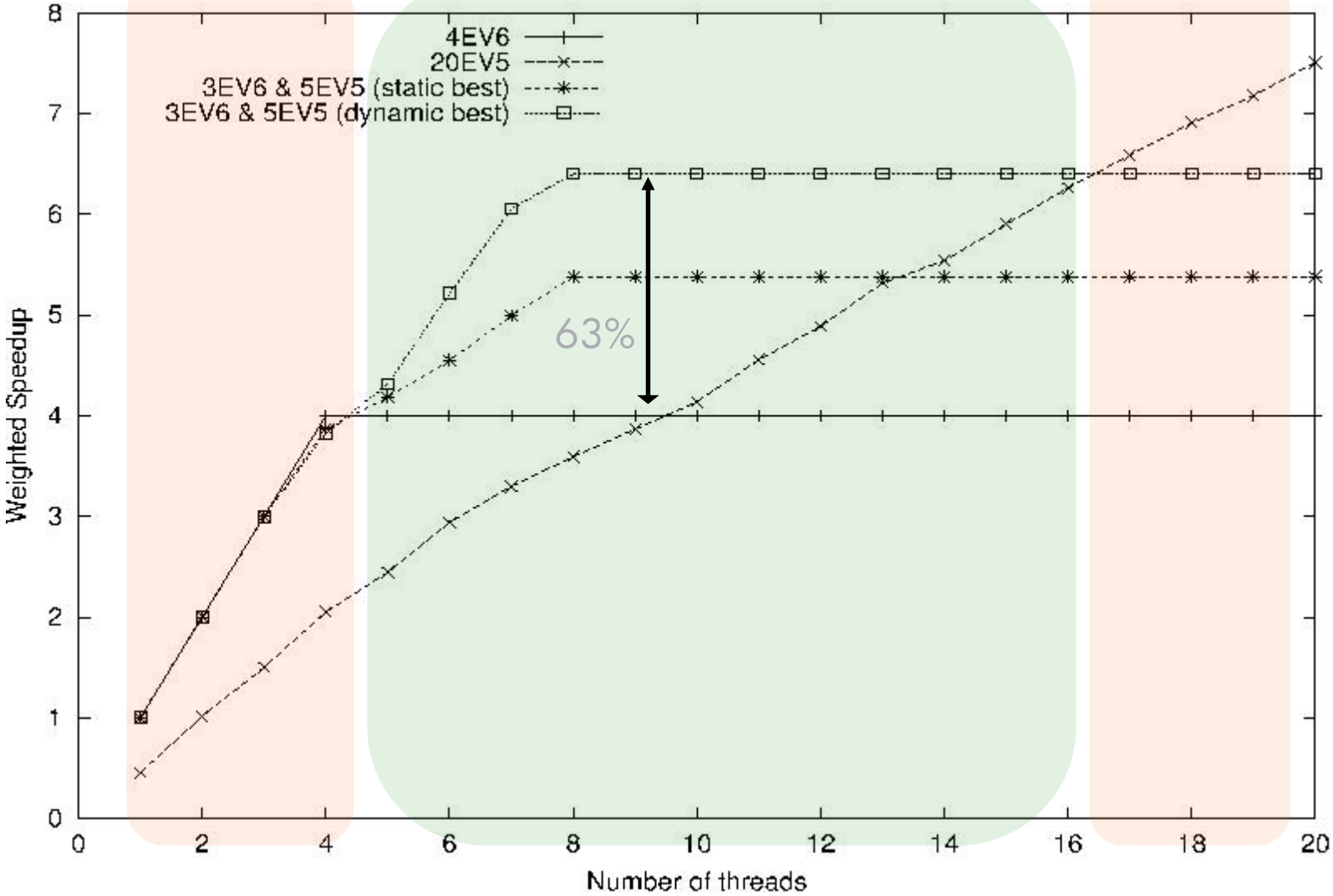


Combine most frequently active blocks into small heterogeneous core instead



Power down large core when not needed

Performance benefits from heterogeneity

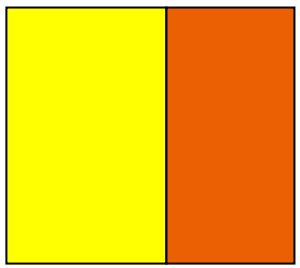


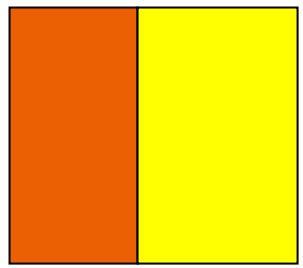
Mitigating Amdahl's Law

- Amdahl's law: Parallel speedups limited by serial portions
- Annavaram et. al., ISCA 2005:
  - Basic idea
    - Use big core for serial portions
    - Use many small cores for parallel portions
  - Prototype built from discrete 4-way SMP
    - Ran one socket at regular voltage, other 3 at low voltage
    - 38% wall clock speedup using fixed power budget

Conjoined Cores

Ideally, provide for peak needs of any single thread without multiplying the cost with the number of cores

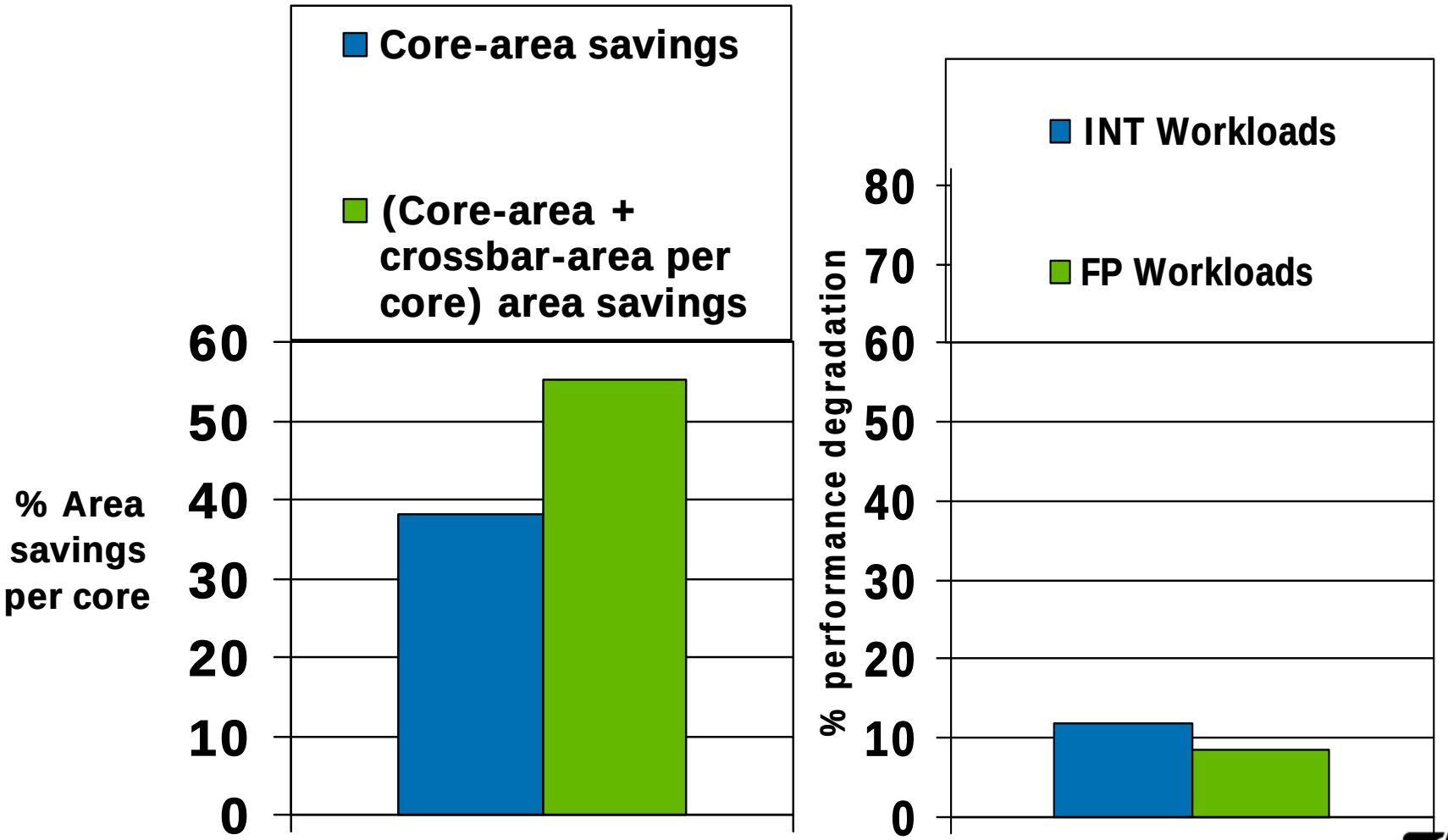




Baseline usage resource

Peak usage resource

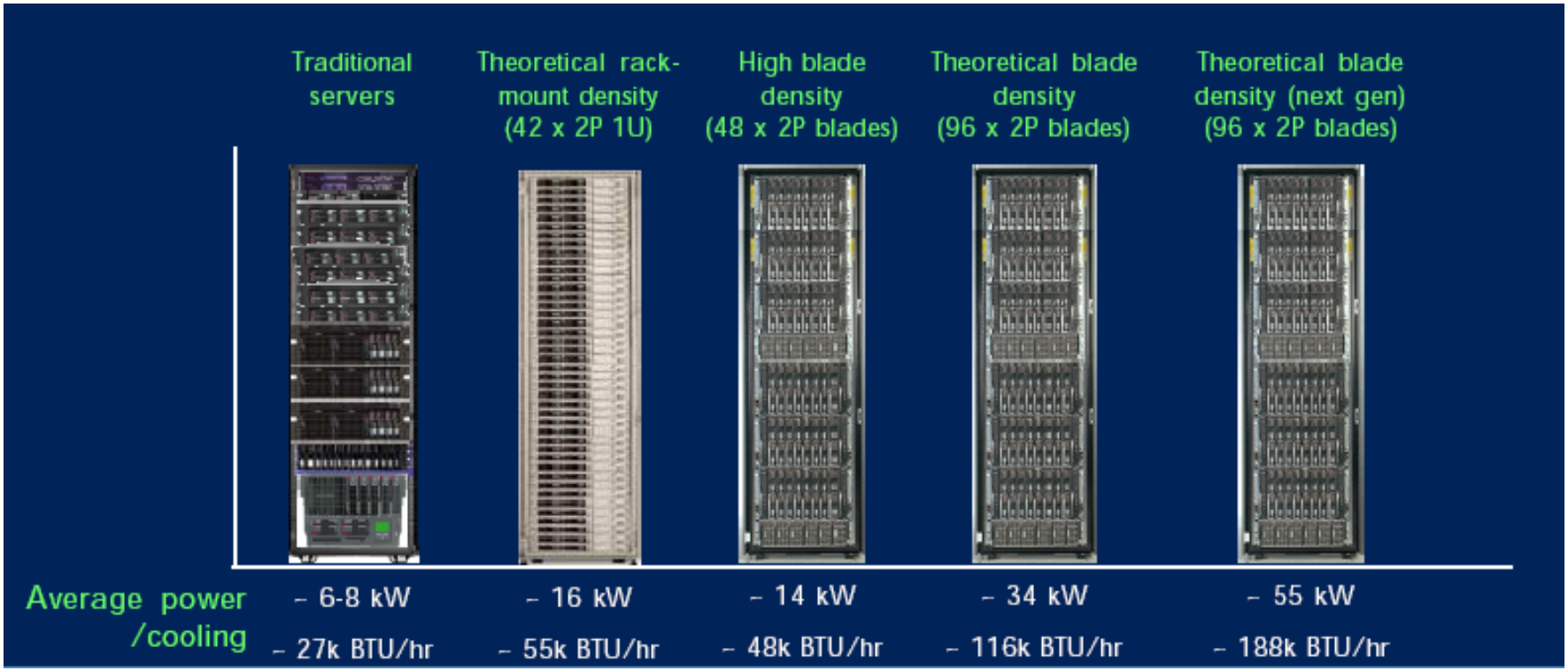
Conjoined-core Architecture Benefits



Performance degradation even less if < 8 threads!

Datacenter Power Optimizations

- A rack of blades can dissipate 30kW or more!



Data Center Power is Front Page News

WSJ.com THE WALL STREET JOURNAL. ONLINE

Free Dow Jones Sites | As of Monday, November 14, 2005

- Home
- News
- Technology
- Markets
- Personal Journal
- Opinion
- Weekend & Leisure

Today's Print Edition
U.S. | Europe | Asia
Past Editions
Features
Portfolio

PAGE ONE

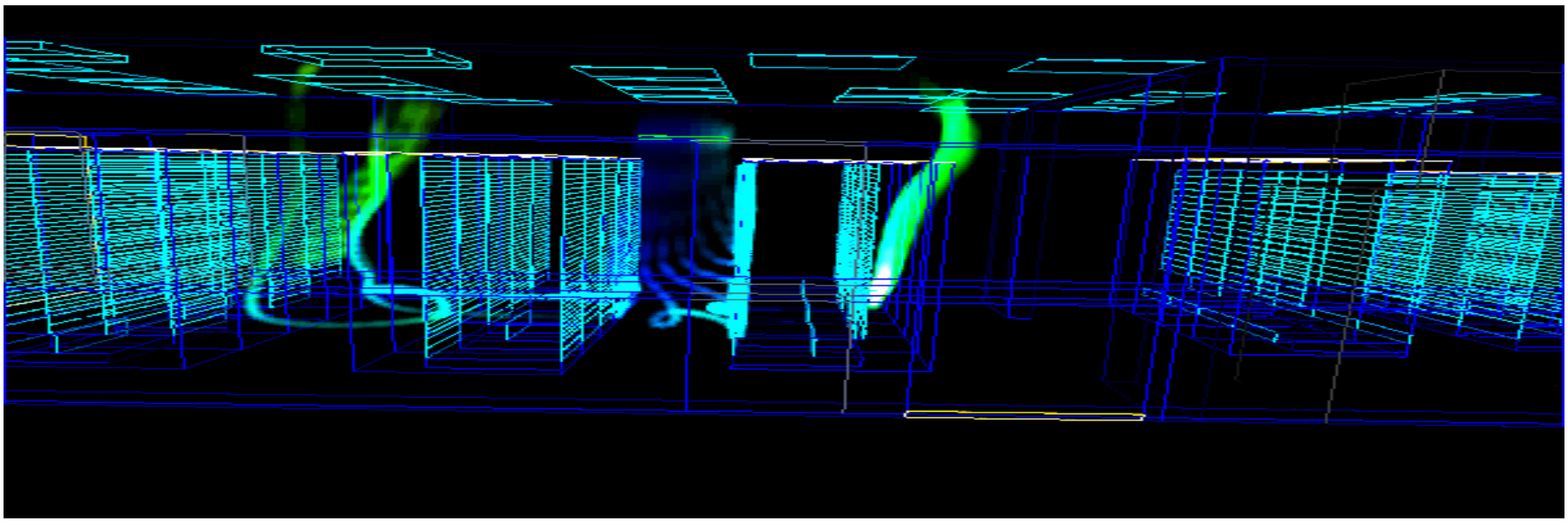
Power-Hungry Computers Put Data Centers in Bind

Newer Hardware Guzzles Electricity and Spews Heat, Requiring Costly Alterations

By DON CLARK
Staff Reporter of THE WALL STREET JOURNAL
November 14, 2005; Page A1

Data center thermal management at HP

- Power density also becoming a significant reliability issue
- Use 3D modeling and measurement to understand thermal characteristics of data centers
  - Saving 25% today
- Exploit this for dynamic resource allocation and provisioning
- Chandrakant Patel et. al.



A Theory on the Evolution of Computer Architecture Ideas

- Conjecture: There is no such thing as a bad or discredited idea in computer architecture, only ideas at the wrong level, place, or time
- Reuse, Recycle
- Evolution vs. Revolution
- Examples:
  - SIMD
  - Dataflow
  - HLL architectures
  - Capabilities
  - Vectors

SIMD

- Efficient way of computing proposed in late '60s
  - 64-PE Illiac-IV operational in 1972
  - Difficult to program

- Intel MMX introduced in 1996
  - Efficient use of larger word sizes for small parallel data
  - Used in libraries or specialized code
  - Only small increase in hardware (<1%)

Dataflow

- Widely researched in late '80s
- Lots of problems:
  - Complicated machines
  - New programming model
- Out-of-order (OOO) execution machines developed in '90s are a limited form of data flow
  - Issue queues issue instructions when operands ready
  - But keeps same instruction set architecture
  - Keeps same programming model
  - Still complex internally

High-Level Language Architectures

- Popular research in 1970's and early 1980's
- "Closing the semantic gap"
- A few attempts to implement in hardware failed
  - Machines interpreted HLLs in hardware
- Now we have Java interpreted in software
  - JIT compilers
  - Portability
  - Modest performance loss doesn't matter for some apps

Capabilities

- Popular research topic in 1970's
- Intel 432 implemented capabilities in hardware
  - Every memory reference
  - Poor performance
- Died with 432
- Security increasingly important
- Capabilities at the file-system level combined with standard memory protection models
  - Much less overhead
- Virtual machine support

Lessons from "Idea Evolution Theory"

- Don't be afraid to look at past ideas that didn't work out as a source of inspiration
- Some ideas make be successful if reinterpreted at a different level, place, or time when they can be made more evolutionary than revolutionary

Conclusions

- The Power Wall is here
  - It is ***the*** wall to worry about
  - It has dramatic implications for the industry
    - From the transistor through the data center
- We need to reclaim past efficiencies
  - Microarchitectural complexity needs to be reduced
- The power wall will usher in the "Golden Age of Parallel Programming"
- Much open research in architecture
- It may be time to reexamine some previously discarded architecture ideas

Thanks

hp